DOCUMENT RESUME

ED 306 281 TM 013 132

AUTHOR Ackerman, Terry A.
TITLE An Explanation of Differential Item Functioning from a Multidimensional Perspective.
SPONS AGENCY Office of Naval Research, Washington, D.C. Personnel and Training Branch.
REPORT NO NR153-531
PUB DATE Apr 88
CONTRACT 1-N00014-85-C-0241
NOTE 30p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; College Entrance Examinations; Higher Education; *Item Analysis; *Latent Trait Theory; Multidimensional Scaling; Standardized Tests; Test Bias; Test Construction; *Test Items
IDENTIFIERS Calibration; *Differential Item Performance; Unidimensionality (Tests)

ABSTRACT

Many researchers have suggested that the cause of differential item functioning (DIF) can in part be due to the misspecification of the supporting trait distribution (STD). This paper demonstrates how a unidimensional item response theory (IRT) calibration of response data, generated from a two-dimensional IRT model, results in DIF when the multidimensional STDs are not equal. How DIF can occur when items measure multiple ability dimensions on which groups have different STDs is illustrated. Generating item parameters used to simulate the multidimensional test were based on the American College Testing (ACT) Mathematics Usage Test administered in February 1983. The calibration sample consisted of 2,000 randomly selected students. With these item estimates as parameters, response data corresponding to three different STD conditions were generated. Results indicate that DIF, created by model misspecification, can be accurately predicted if the multidimensional IRT item parameters and the STD for the groups of interest are known. Implications and directions for future research are discussed. Five tables and four graphs present study data. (Author/SLD)

An Explanation of Differential Item Functioning from a Multidimensional Perspective

Terry A. Ackerman

The American College Testing Program

Paper presented at the 1988 AERA Annual Meeting, New Orleans, April 6, 1988. This research was supported by Contract Number 1-N00014-85-C-0241, NR153-531 from the Personnel and Training Research Programs of the Office of Naval Research.

Running Head: Differential Item Functioning

Abstract

Many researchers have suggested that the cause of differential item functioning (DIF) can in part be due to the misspecification of the supporting trait distribution (STD). This paper demonstrates how a unidimensional IRT calibration of response data, generated from a two-dimensional IRT model, results in DIF when the multidimensional STDs are not equal. Results indicate that DIF, created by model misspecification can be accurately predicted if the multidimensional IRT item parameters and the STD for the groups of interest are known. Implications and directions for future research are discussed.

An Explanation of Differential Item Functioning from a Multidimensional Perspective

Introduction

It is the purpose of most standardized tests to distinguish between levels of traits for individuals or groups of individuals. However, because the underlying or supporting trait distribution (STD) of a group of examinees is not directly observable, researchers have developed models which can be used to describe the relationship between an observable response and the latent STD which generated the response. By using such models, researchers try to determine the differences not only between the observable responses but more importantly the STDs. A problem arises when there is not a direct or one-to-one relationship between the observable response and the quantified underlying ability which produced it. Model misspecification is a problem that researchers have been faced with for years especially when one tries to model cognitive process (cf. Traub, 1983).

For pragmatic reasons, educators, although aware of these complex cognitive processes, base decisions concerning individual abilities or group ability distributions on only the observable responses. Many standardized achievement tests report a single content score, suggesting that all of the items that produce such a score are measuring only that content. For example, when a single math score representing the performance on a given set of math items is reported, it is assumed that all of the items measure only the reported math ability. However, should some of the math items require more of another ability, such as reading, for a correct response then the reported score cannot fairly be termed a math score. Also, and more importantly, if the reading component is substantial, then the test will favor the better readers, regardless of whether or not they have identical math ability!

Over the past several years there has been a proliferation of research examining how standardized tests measure individuals to determine if certain measures of cognition are biased or favor certain individuals or groups of individuals. A test which is used for determining college placement or making scholarship decisions could have a profound effect if examinee groups, which have identical STD on the reported abilities, perform differently because extraneous, nonreported skills are required for a correct response.

Many researchers (Lord, 1980; Linn & Harnisch, 1981; Traub, 1983, Wang 1986) have suggested that one of the major causes of differential item functioning (DIF) is that the "biased" items are measuring abilities other than those of the reported test score. It is the purpose of this paper to illustrate, within a multidimensional item response theory (MIRT) framework, how different STD produce DIF, when the complete ability space which is required for a correct response is misspecified. Examples will be shown that will demonstrate how unidimensional IRT estimates of items that require two-dimensional abilities for a correct response, will result in biased estimates of ability when responses originate from disparate two-dimensional abilities.

Background

The work of Reckase (1986) in formally defining MIRT item characteristics provides an appropriate framework with which hypothetical multidimensional abilities and items can be easily specified and item response data can be subsequently generated. Reckase (1986) defined the unidimensional IRT counterparts of difficulty and discrimination for the multidimensional M2PL model, given by

$$P(\underline{X}_{ij} = 1 | a_i, \underline{d}_i, \theta_j) = \frac{\exp(a_i' \theta_j + \underline{d}_i)}{1 + \exp(a_i' \theta_j + \underline{d}_i)} \quad (1)$$

where $\underline{X}_{ij}$ is the score (0, 1) on item i by person j,

a_i is the vector of item discrimination parameters

$\underline{d}_i$ is a scalar parameter related to the difficulty of the item

and θ_j is the vector of ability parameters for person j.

In a two-dimensional latent ability space (e.g., math and verbal ability dimensions), the a_i vector designates the degree to which an item distinguishes between individuals on both abilities. Thus, if an item had the parameters $\underline{a}_1 = \underline{a}_2$ it would be distinguishing between individuals on both dimensions equally well. However, if $\underline{a}_1 = 0$ and $\underline{a}_2 = 1$, an item would be discriminating only along the θ_2 dimension.

Using the notation of Reckase (1985) an item j, which requires two abilities for a correct response can be represented in the two-dimensional ability plane as a vector of length $\underline{D}$ in the direction α_j, where

$$D_j = \frac{-d_j}{(a_{1j}^2 + a_{2j}^2)^{\frac{1}{2}}} \quad (2)$$

and

$$\alpha_j = \arccos \frac{a_{ij}}{(a_{1j}^2 + a_{2j}^2)^{\frac{1}{2}}} . \quad (3)$$

Because the discrimination parameters can never be negative, the vectors, which start at $\theta_1 = 0$, $\theta_2 = 0$, lie only in the third quadrant when D_j is negative, (representing easy items) or in the first quadrant when D_j is positive (representing difficult items.)

Figure 1 illustrates an item vector whose M2PL parameters are given as $d = 2.00$, $a_1 = 1.00$, $a_2 = 1.00$. Also illustrated in Figure 1 are the lines of equiprobability. Notice that for the M2PL model these lines will always be parallel which is indicative of the compensatory nature of this model. (That is, high ability on one dimension will compensate to some extent for a low ability on a second dimension.)

Insert Figure 1 about here

Because the vector is describing the direction and distance to the line of inflection on the item's response surface, it can also be seen in Figure 1 that the $p = .5$ probability line runs orthogonal to the tip of the item vector. Thus, if the STD of a group is known then the proportion of correct responses for a group on an item can be easily estimated.

For example, consider two Groups A and B having the STDs as shown in Figure 2. Assume both groups have the same mean ability for mathematics (dimension 2), but because of instructional differences Group B has a higher mean reading ability (dimension 1). That is, $\bar{\theta}_2 = 1.2$ for both groups, but $\bar{\theta}_1$ equals 0.0 and 2.3 for Groups A and B, respectively. Assume further that each group is given a test, purported to be a mathematics test, which consists of three items. The item 1 vector, as drawn in Figure 2, measures only mathematics ability; item 2 measures both math and reading equally, and item 3, whose vector is orthogonal to item 1 measures solely reading ability.

Insert Figure 2 about here

The success of the two groups on each item can be roughly determined by examing the p = .5 equiprobability line in relationship to each group's ability centroid. Both groups will perform equally well on the math item, (Item 1) but Group B should easily outperform Group A on items 2 and 3. Thus even though the two groups have identical math ability distributions, this mathematics test will favor Group B (i.e., Group B would have a higher expected raw score) because some of the items are measuring an ability (reading) which is not being considered in the reported mathematics score, but is essential to produce a correct response for the majority of test items.

Whereas Reckase's work is more from a geometric perspective, other researchers have approached the relationship between multidimensional and unidimensional IRT models from a more analytic framework. Wang (1986) determined explicit algebraic relationships between unidimensional estimates and the true multidimensional parameters for the case in which the underlying response process is modeled by the M2PL MIRT response model and the unidimensional model is the two parameter logistic (2PL) model in which the probability of a correct response is given as:

$$P(X = 1|\theta) = \left[1.0 + \exp\left(-1.7a(\theta - b)\right)\right]^{-1} \qquad (4)$$

where a and b are the unidimensional discrimination and difficulty parameters and θ is the unidimensional latent ability measure. Using the analytical results for unidimensional approximation to a multidimensional data matrix,

Wang concluded that the unidimensional estimates of the item parameters are obtained with reference to a weighted composite of the underlying latent traits. The weights are primarily a function of the discrimination vectors for the items, the correlations among the latent traits and, to a lesser extent, the difficulty parameters of the items.

Specifically, for a group g whose STD can be described as having a diagonal variance-covariance structure $\underset{\sim}{\Omega}_g$ and mean ability vector, μ, the unidimensional 2PL IRT item parameters for item j can be approximated by

$$\hat{a}_j = \underset{\sim}{a}_j' \underset{\sim}{W}_1 / \sqrt{2.89 + \underset{\sim}{a}_j' \underset{\sim}{W}_2 \underset{\sim}{W}_2' \underset{\sim}{a}_j} \qquad (5)$$

$$\hat{b}_j = (d_j - \underset{\sim}{a}_j' \underset{\sim}{\mu}) / \underset{\sim}{a}_j' \underset{\sim}{W}_1 \qquad (6)$$

in which $\underset{\sim}{a}_j$ is the discrimination vector for the M2PL model

d_j is the difficulty parameter for the M2PL model

$\underset{\sim}{W}_1$ and $\underset{\sim}{W}_2$ are the first and second standardized eigenvectors of the matrix L'A'AL, where,

A is the matrix of discrimination parameters for all the items in the test and $L'L = \Omega$.

Thus, when the first two moments of a 2-dimensional STD are known, as well as the 2-dimensional item parameters, the corresponding 2PL IRT unidimensional item parameter estimates (as computed by a calibration program such as LOGIST (Wingersky, Barton, & Lord, 1982) can be easily approximated.

It is the purpose of this paper to provide a simple illustration of how DIF can occur when items measure multiple ability dimensions on which groups have different STD. Specifically, the paper will demonstrate how DIF can be

predicted using the formulation of Reckase (1985) and Wang (1986) when the multidimensional item parameters and multidimensional STDs for the groups of interest are known.

Methodology

The generating item parameters used to simulate the multidimensional test were based on the M2PL estimates of the ACT Mathematics Usage Test which was administered in February 1983. These item parameter estimates were determined using the computer program MIRTE (Carlson, 1987). This program estimates the M2PL parameters using a joint maximum likelihood procedure. The calibration sample consisted of 2000 randomly selected students. These item parameter estimates were used as a representative sample of a multidimensional standardized mathematics test.

Using these item estimates as parameters, response data corresponding to three different STD conditions were generated. In all, 2000 randomly created subjects were generated for each group according to the group's specified STD. The characteristics of each group are described in Table 1.

Insert Table 1 about here

The differences in the STD are purely for illustrative purposes although such differences could conceivably occur through instructional differences. Group A, the base or reference group, has a STD distribution represented by a bivariate normal distribution centered at $(\bar{\theta}_1, \bar{\theta}_2) = (0, 0)$. Groups B and C,

the two primary or focal groups of interest have mean abilities centered (1.0, -0.5) and (-0.5, 1.0). Group B is about 1.7 times more variable along the first dimension than the second dimension and Group C is about 1.7 times more variable along the second dimension than the first dimension.

Using the mean and variance of each group in concert with the generating item parameters, unidimensional item parameters corresponding to the 2PL IRT model were computed using equations 4 and 5. Using two different methods, the performance of each focal group was predicted to be better, equal to, or worse than the performance of the preference group for each of the 40 items. One set of predictions involved identifying the direction and length of each item's 2-D vector in relationship to each group's STD. For Group B, items whose angles were less than 63° were predicted to favor Group B, and items whose angles were greater than 63° were predicted to favor Group A. An item whose direction was 63° was predicted to be equally difficult for both groups. The angle of 63° was chosen because it represents the composite direction perpendicular to the line connecting the mean ability centroids for Groups A and B. Because the equiprobability lines of the M2PL model are parallel and orthogonal to the composite of abilities being measured, shifting the mean centroid of a group orthogonally to an item vector is equivalent to moving the centroid along a line of equiprobability. Thus, there should be no difference in group performance on that item. For Group C the same procedure was followed but with an angle of 37°.

The second set of predictions were based upon the 2PL model's logit, $\hat{a}\theta - \hat{a}\hat{b}$ where $\hat{a}$ and $\hat{b}$ are the analytically computed item parameters and θ, the ability parameter, was set equal to 0.0. It can be seen from Equation 6 that the greater the value of the $-\hat{a}\,\hat{b}$ term, the greater the probability of correct response. Thus, by comparing the logit

(where $\theta = 0$) for Group B and C to the logit of Group A, (i.e., is the logit of Group B less than, equal to, or greater than the logit of Group A for item j) predictions about the focal groups' relative probability of correct response were made for each item.

After predictions were made, the generated response data for Set A were calibrated using LOGIST (Wingersky, Barton, & Lord, 1982), and 2PL IRT item and ability parameters were estimated. Sets B and C were also calibrated using LOGIST. However, in each of these runs the item parameters were fixed to those values estimated for Set A and only abilities were estimated. This procedure assured that abilities for each focal group would be placed on the same scale as Group A.

As a measure of model fit and DIF, the Linn-Harnisch Z statistic was computed for each item, for each group. The statistic is computed as

$$Z_i = \frac{1}{N} \sum_{j=1}^{n} \frac{u_{ij} - P_{ij}}{\sqrt{P_{ij}(1 - P_{ij})}} \qquad (7)$$

where u_{ij} is the 1 (correct) or 0 (incorrect) response of person j to item i and P_{ij} is the 2PL IRT model given in equation 6 using Set A's item parameter estimates. If, for example, Group B had a higher probability of correct response than Group A, Z_i for Group B would be greater than Z_i for Group A. If Group B had a lower probability of correct response than Set A, Z_{iB} would be less than Z_{iA}. For each item the observed DIF computed using the Z statistic was compared to the predicted performance of each focal group. The proportion of correct predictions was then determined.

Results

The M2PL item parameter estimates from the program MIRTE (Carlson, 1987) are presented in Table 2. The direction or two-dimensional ability composite that is being measured ranges from measuring only the first dimension (items 10, 11, and 18) to measuring totally the second dimension (item 34). The AAP Math Usage Test is constructed so that as an examinee proceeds through the test, the items become increasingly more difficult. Interestingly, this shift in difficulty is reflected by a shift from measuring primarily the θ_1 ability to measuring primarily the θ_2 ability as the item number increases.

Insert Table 2 about here

Substantively these items were reviewed to determine if the two dimensions could be identified. Those items which primarily measured the first dimension were classified as Arithmetic and Algebraic Reasoning or AAR items. These items were basically story problems in which the examinee had to read through a two to three sentence passage before responding to the item. Thus it was felt that this passage measured the verbal loading of an item. The second dimension was classified as a computation dimension, since the items which discriminated best along this dimension were found to be Intermediate Algebra (IA) and Advanced Topic (AT) items. These items had very short item stems and required the examinee to perform more computational type operations steps to solve the problem.

The fit of the Math Usage data to the M2PL model was checked by examining the distribution of residual covariances as output by the program MIRTE. This distribution was highly skewed with most residuals in the range from 0.0 to .011. A two-dimensional adaptation of Yen's (1983) (See McKinley and Mills, 1985) chi square goodness-of-fit statistic also showed that none of the items had a significant chi-square value indicating no lack of model fit.

The mean ability vector and covariance structures of the three generated ability distributions are listed in the middle of Table 1. The IMSL subroutine GGNSM was used to generate each distribution and appeared to produce the hypothetical distributions quite well. The first three eigenvalues of a principal component analysis of the tetrachoric correlation matrix for each response set are shown at the bottom of Table 1. These values helped verify that the data were multidimensional (see Reckase, 1979).

The raw score means were 17.55, 21.54 and 18.61 for Groups A, B, and C, respectively. The standard deviations of the raw score distributions were all about equal to 1.2. These results are interesting because they indicate that Group B performed better than Groups A and C even though A and C have a higher mean computational ability. Because more of the items have a verbal loading than a computational loading (i.e., measure better along the θ_1 dimension) the group which has the greatest "reading ability" will obtain the highest mean raw score.

One way to confirm the mean raw scores for each group is to sketch the three STD's upon a contour plot of the expected raw score for all individuals in the two-dimensional ability plane. This plot is displayed in Figure 3. Each line can be considered an "equi-true score". For example, each (θ_1, θ_2) which lies on the line denoted 36, has an expected true score of 36 on the 40-item test. It can be seen in Figure 3, that the ability

centroids for each group lie very closely to the "equi-true score" line that matches their computed raw score mean based upon the generated data. Thus Group A centroid lies between the equi-true score lines representing expected values of 17 and 18. Similar results can be seen for Groups B and C.

Insert Figure 3 about here

Another way to illustrate the difference in item performance of Groups B and C relative to Group A is a plot of the p-value difference for each focal group for each item. This plot, shown in Figure 4, illustrates how, on the majority of items, Group B outperformed Group A. Only on the later items (29-40, except #32) did Group C have a larger percentage of correct responses than either Groups A or B.

Insert Figure 4 about here

The analytical 2PL item parameter estimates based upon the work of Wang (1986) are shown in Table 3. Both the $\hat{a}$ and $\hat{b}$ for Sets B and C were rescaled and placed on the same scale as the Set A estimates. If there was no DIF, these parameter estimates would be identical for each group.

Insert Table 3 about here

Depending on the multidimensional composite an item is measuring, the item parameter estimates vary considerably from each other. Items which primarily load on the verbal dimension (θ_1) such as 10 and 18 discriminate better for Group B than either Set B or especially Set C. The reverse is true for items which measure primarily the computation dimension (θ_2). Items such as 34 and 35 discriminate much better for Group C than either Groups A or B.

Large differences also appear between the analytical difficulty parameter estimates for each of the three groups. Items 10 and 11 would be considered to be quite easy for Group B ($\hat{b}$ = -1.24 and -1.08, respectively), and about average for Group A ($\hat{b}$ = -.23 and -.04, respectively.) Some items which are moderate in difficulty for Set C (such as item 34 ($\hat{b}$ = -.01) were difficult for Set A ($\hat{b}$ = 1.79) and extremely difficult for Set B ($\hat{b}$ = 5.00).

The LOGIST item parameter estimates are also displayed in Table 3 in columns 5 and 9. These tend to be quite similar to the Set A analytical estimates even though the analytical estimates were not rescaled to the LOGIST item parameters to account for differences due to sampling error. The LOGIST calibration run converged on the 2PL solution in 15 stages.

One set of DIF predictions was based upon the analytical item estimates. The logit was ($\hat{a}\theta - \hat{a}\hat{b}$, where $\theta = 0$) computed for each item for each group. These are shown in columns 2, 3 and 4 of Table 4. The larger the logit, the greater the group's probability of correct response. Again, noticeable differences appear for each group.

Insert Table 4 about here

Using the Set A item parameters, the abilities for Set B and Set C were computed, also using LOGIST. The ability estimates were then used to compute the Linn-Harnisch Z statistic for each of the three groups. These results are reported in the second half of Table 4.

If the 2PL model fit the data well, an item's Z statistic which is summed over all people in the respective group, should be zero. It appears, upon examining the Z statistics for Set A, that LOGIST fit the generated response data extremely well. However, the Z statistics for Groups B and C are quite different, usually having opposite signs.

The "hit" rates (percent of correct predictions) were determined for each group. These results are displayed in Table 5. The predictions for Set B were not as good as those made for Set C. Using the "logit method," the percent of item performance predicted correctly was 70% and 100% for Set B and C, respectively. The "item vector method" provided a 73% and 95% hit rate for Set B and C, respectively.

Insert Table 5 about here

Discussion

This study illustrates how DIF can occur when there is a misspecification of the latent ability space. Based upon the examples provided in this study,

a word of caution should be extended to standardized test authors that the relationship between reported scores and underlying STDs of various examinee groups might need to be studied. To make a standardized test not only fair for all examinees, but also informative, scores and the abilities they measure, need to be clearly and accurately stated. Hopefully, multidimensional IRT will provide new methodology that not only can detect DIF, but can provide some substantive support about why groups perform the way they do. The test creation process could be improved if the relationship between item type and the STD of different examinee groups were more clearly understood and shared.

It needs to be reiterated that the example presented here was designed for illustrations purposes and may not be totally realistic. Before generalizing to other situations several questions pertaining to this study need to be further explored. First of all, how realistic were the three generating STDs? It may be argued that all cognitive abilities are correlated to some degree, and that this factor should be taken into account. How discrepant, for whatever reasons, are STD between groups of interest? How realistic is the compensatory M2PL model? These questions need to be further explored in future studies.

References

Carlson, J. E. (1987). Multidimensional item response theory estimation: a computer program. (Research Report 87-19). Iowa City, IA: The American College Testing Program.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. Journal of Educational Measurement, 18, 109-118.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. Applied Psychological Measurement, 9, 49-58.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4, 207-230.

Reckase, M. D. (1985, April). The difficulty of test items that measure more than one ability. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Reckase, M. D. (1986, April). The discriminating power of items that measure more than one dimension. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), Applications of item response theory. Vancover: Educational Research Institute of British Columbia, 57-70.

Wang, M. (April 1986). Fitting a unidimensional model to multidimensional item response data. Paper presented at the ONR contractors conference. Gatlinburg, TN: To appear as ONR Technical Report 87-1. University of Iowa.

Wingersky, M. S., Barton, M. A., and Lord, F. M. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Table 1

Descriptive statistics of the generating and obtained STDs for the three hypothetical groups

	Group		
	A (reference)	B (focal)	C (focal)
	Hypothetical Distribution		
Mean Ability	$\bar{\theta}_1 = 0$ $\bar{\theta}_2 = 0$	$\bar{\theta}_1 = 1.0$ $\bar{\theta}_2 = -.5$	$\bar{\theta}_1 = -.5$ $\bar{\theta}_2 = 1.0$
Covariance Structure	$\begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix}$	$\begin{vmatrix} 1.5 & 0 \\ 0 & .5 \end{vmatrix}$	$\begin{vmatrix} .5 & 0 \\ 0 & 1.5 \end{vmatrix}$
	Generated Distribution		
Mean Ability	$\bar{\theta}_1 = .014$ $\bar{\theta}_2 = -.020$	$\bar{\theta}_1 = 1.017$ $\bar{\theta}_2 = .974$	$\bar{\theta}_1 = -.489$ $\bar{\theta}_2 = .974$
Covariance Structure	$\begin{vmatrix} 1.02 & .02 \\ .02 & 1.08 \end{vmatrix}$	$\begin{vmatrix} 1.54 & .02 \\ .02 & .54 \end{vmatrix}$	$\begin{vmatrix} .51 & .02 \\ .02 & 1.61 \end{vmatrix}$
	Eigenvalues		
Principal Component			
1	8.999	9.399	8.162
2	2.732	2.092	2.637
3	1.040	1.060	1.040

Table 2

Multidimensional Parameter Estimates, Directions and Distances for the Items in the ACT Assessment Mathematics Usage Test

Item	$\hat{a}_{i1}$	$\hat{a}_{i2}$	$\hat{d}_i$	α_i	D_i
1	1.81	.86	1.46	25	-.73
2	1.22	.02	.17	1	-.14
3	1.57	.36	.67	13	-.42
4	.71	.53	.44	37	-.50
5	.86	.19	.10	12	-.11
6	1.72	.18	.44	6	-.25
7	1.86	.29	.38	9	-.20
8	1.33	.34	.69	14	-.50
9	1.19	1.57	.17	53	-.09
10	2.00	.00	.38	0	-.19
11	.87	.00	.03	0	-.03
12	2.00	.98	.91	26	-.41
13	1.00	.89	-.49	42	.37
14	1.22	.14	.54	7	-.44
15	1.27	.47	.29	20	-.21
16	1.35	1.15	-.21	40	.12
17	1.06	.45	.08	23	-.07
18	1.92	.00	.12	0	-.06
19	.96	.22	-.30	13	.30
20	1.20	.12	-.28	6	.23
21	1.41	.04	-.21	2	.15
22	1.54	1.79	.02	49	-.01
23	.54	.23	-.69	23	1.18
24	1.53	.48	-.83	17	.52
25	.72	.55	-.56	37	.62
26	.51	.65	-.49	52	.59
27	1.66	1.72	-.38	46	.16
28	.69	.19	-.68	15	.95
29	.88	1.12	-.91	52	.64
30	.68	1.21	-1.08	61	.78
31	.24	1.14	-.95	78	.82
32	.51	1.21	-1.00	67	.76
33	.76	.59	-.96	38	1.00
34	.01	1.94	-1.92	90	.99
35	.39	1.77	-1.57	78	.87
36	.76	.99	-1.36	52	1.09
37	.49	1.10	-.81	66	.67
38	.29	1.10	-.99	75	.87
39	.48	1.00	-1.56	64	1.41
40	.42	.75	-1.61	61	1.87

Table 3

Analytical and LOGIST Calibrated Parameter Estimates for Each Item By Group

Item	$\hat{a}$ Group A	$\hat{a}$ Group B	$\hat{a}$ Group C	LOGIST	$\hat{b}$ Group A	$\hat{b}$ Group B	$\hat{b}$ Group C	LOGIST
1	1.15	1.15	.60	1.33	-.74	-1.43	-1.05	-.72
2	.57	.68	.18	.56	-.17	-1.18	1.12	-.16
3	.84	.94	.37	.87	-.44	-1.28	-.31	-.42
4	.52	.48	.40	.50	-.50	-1.06	-.86	-.51
5	.48	.52	.23	.53	-.12	-.99	.32	-.18
6	.82	.98	.29	.87	-.29	-1.22	.36	-.23
7	.92	1.08	.34	1.07	-.22	-1.12	.32	-.18
8	.74	.81	.35	.80	-.53	-1.35	-.51	-.55
9	1.02	.78	1.01	1.08	-.09	-.36	-.62	-.07
10	.83	1.06	.23	.93	-.23	-1.24	1.06	-.22
11	.41	.48	.13	.46	-.04	-1.08	1.60	-.11
12	1.28	1.27	.65	1.52	-.41	-1.10	-.59	-.36
13	.77	.68	.62	.84	.37	-0.05	.09	.36
14	.61	.70	.24	.74	-.49	-1.40	-.14	-.42
15	.76	.79	.41	.77	-.22	-.98	-.15	-.22
16	1.03	.88	.77	1.21	.12	-.35	-.18	.10
17	.66	.67	.39	.64	-.07	-.80	0.00	-.02
18	.80	1.03	.22	.84	-.08	-1.11	1.50	-.11
19	.53	.58	.26	.58	.33	-.56	1.14	.31
20	.60	.69	.23	.68	.26	-.73	1.63	.22
21	.65	.78	.21	.68	.18	-.87	1.94	.08
22	1.25	.94	1.11	1.40	-.01	-.33	-.48	.03
23	.34	.34	.22	.37	1.19	.45	1.93	1.01
24	.88	.94	.42	.94	.54	-.29	1.23	.52
25	.53	.49	.41	.59	.62	.14	.50	.63
26	.46	.38	.45	.47	.63	.45	.12	.63
27	1.31	1.02	1.06	1.55	.16	-.20	-.24	.14
28	.40	.42	.21	.40	1.00	.12	2.18	.89
29	.77	.62	.75	.83	.67	.51	.17	.64
30	.68	.51	.79	.74	.88	1.00	.16	.84
31	.44	.28	.68	.47	1.15	2.28	-.06	1.12
32	.59	.43	.77	.63	.92	1.30	.03	.91
33	.56	.52	.44	.53	1.00	.55	.95	1.11
34	.46	.23	1.04	.50	1.79	5.00	-.01	1.64
35	.61	.38	1.06	.59	1.21	2.32	-.00	1.28
36	.67	.55	.67	.77	1.15	1.08	.64	1.03
37	.56	.41	.70	.56	.80	1.10	-.04	.81
38	.45	.31	.67	.44	1.17	2.09	.03	1.27
39	.53	.40	.64	.56	1.64	2.10	.73	1.65
40	.44	.35	.49	.48	2.11	2.52	1.27	2.00

Table 4

The Logit and Linn-Harnisch Z Values for Each Item By Group

Item	LOGIT GROUP			LINN-HARNISCH Z GROUP		
	A	B	C	Z_A	Z_B	Z_C
1	.85	1.65	.63	-.00	.04	-.06
2	.09	.80	-.20	-.00	.23	-.32
3	.37	1.21	.11	.00	.16	-.23
4	.26	.51	.35	.00	.03	.06
5	.06	.51	-.08	-.01	.12	-.17
6	.24	1.20	-.10	-.00	.21	-.37
7	.20	1.22	-.11	-.00	.24	-.34
8	.39	1.09	.18	-.00	.12	-.22
9	.09	.28	.63	.00	-.15	.21
10	.19	1.32	-.24	-.00	.24	-.47
11	.02	.52	-.21	-.00	.18	-.25
12	.53	1.39	.38	.00	.06	-.07
13	-.29	.04	-.05	.00	-.06	.10
14	.30	.99	.03	.00	.16	-.26
15	.17	.78	.06	-.00	.14	-.13
16	-.12	.31	.14	.01	-.06	.09
17	.05	.54	.00	-.00	.13	-.09
18	.06	1.14	-.34	-.00	.25	-.49
19	-.17	.32	-.29	.00	.17	-.16
20	-.16	.50	-.37	-.00	.18	-.28
21	-.11	.68	-.41	.00	.25	-.39
22	.01	.31	.53	.02	-.11	.28
23	-.41	-.16	-.42	.00	.07	-.04
24	-.47	.27	-.52	.00	.19	-.20
25	-.33	-.07	-.21	.00	.00	.04
26	-.29	-.17	-.06	.00	-.07	.13
27	-.21	.20	.25	-.00	-.11	.23
28	-.40	-.05	-.46	-.00	.10	-.10
29	-.52	-.31	-.13	-.00	-.17	.21
30	-.60	-.52	-.12	.01	-.21	.28
31	-.50	-.65	.04	.00	-.31	.40
32	-.54	-.55	-.03	.00	-.26	.37
33	-.56	-.29	-.42	-.00	.05	.05
34	-.82	-1.15	.01	.00	-.44	.62
35	-.74	-.89	.00	.00	-.37	.53
36	-.78	-.59	-.43	.01	-.13	.21
37	-.45	-.45	.03	.00	-.23	.32
38	-.53	-.65	-.02	-.00	-.25	.37
39	-.87	-.84	-.47	-.00	-.18	.28
40	-.92	-.87	-.63	.00	-.12	.22

Table 5

Tables Illustrating the Predictions of Each Focal Group Compared to Group A for Each Method of Prediction

Prediction Type

LOGIT

SET B		Observed			
		< A	=A	>A	
Predicted	< A	5	0	0	
	= A	1	0	0	
	> A	10	1	23	

Hit rate: 70%

ITEM VECTOR

		Observed		
		<A	=A	>A
Predicted	< A	7	0	0
	= A	0	0	0
	> A	10	1	22

Hit rate: 73%

Prediction Type

LOGIT

SET C		Observed			
		<A	=A	>A	
Predicted	< A	20	0	0	
	= A	0	0	0	
	> A	0	0	20	

Hit rate: 100%

ITEM VECTOR

		Observed		
		<A	=A	>A
Predicted	< A	21	0	0
	= A	0	1	1
	> A	0	0	17

Hit rate: 95%

Figure Captions

Figure 1. The item vector and equiprobability lines of an M2PL item with A1 = 1.00, A2 = 1.00 and d = 2.00.

Figure 2. The STD for hypothetical Groups A and B and the p = .5 probability lines for the three item math test.

Figure 3. The STD for the generated groups A, B and C and the expected true score lines for the 40 item multidimensional math test.

Figure 4. A plot of the p-value differences between the focal groups and the reference group for each of the 40 math items.

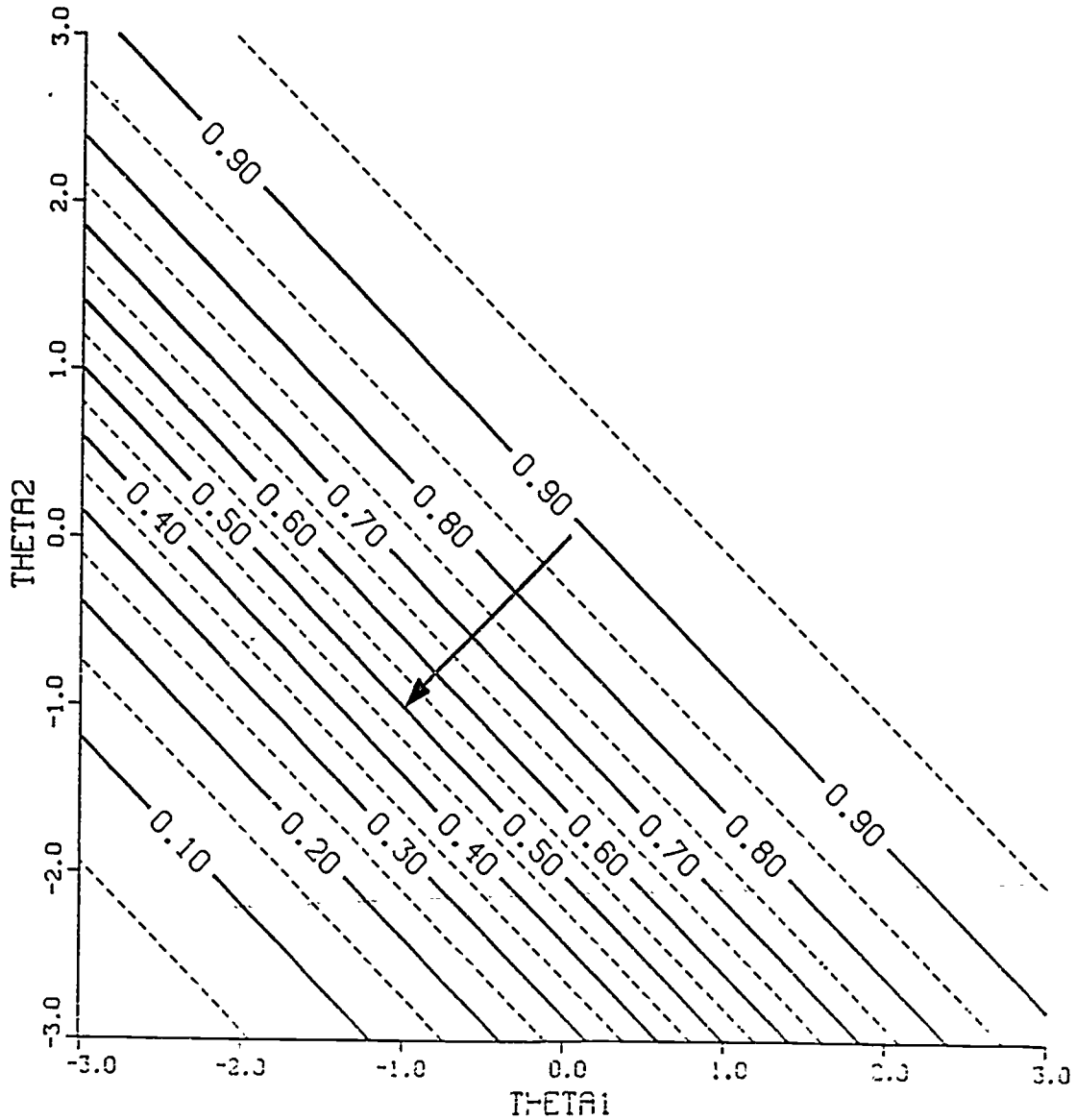
THETA2
THETA1
3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10

FIGURE 1

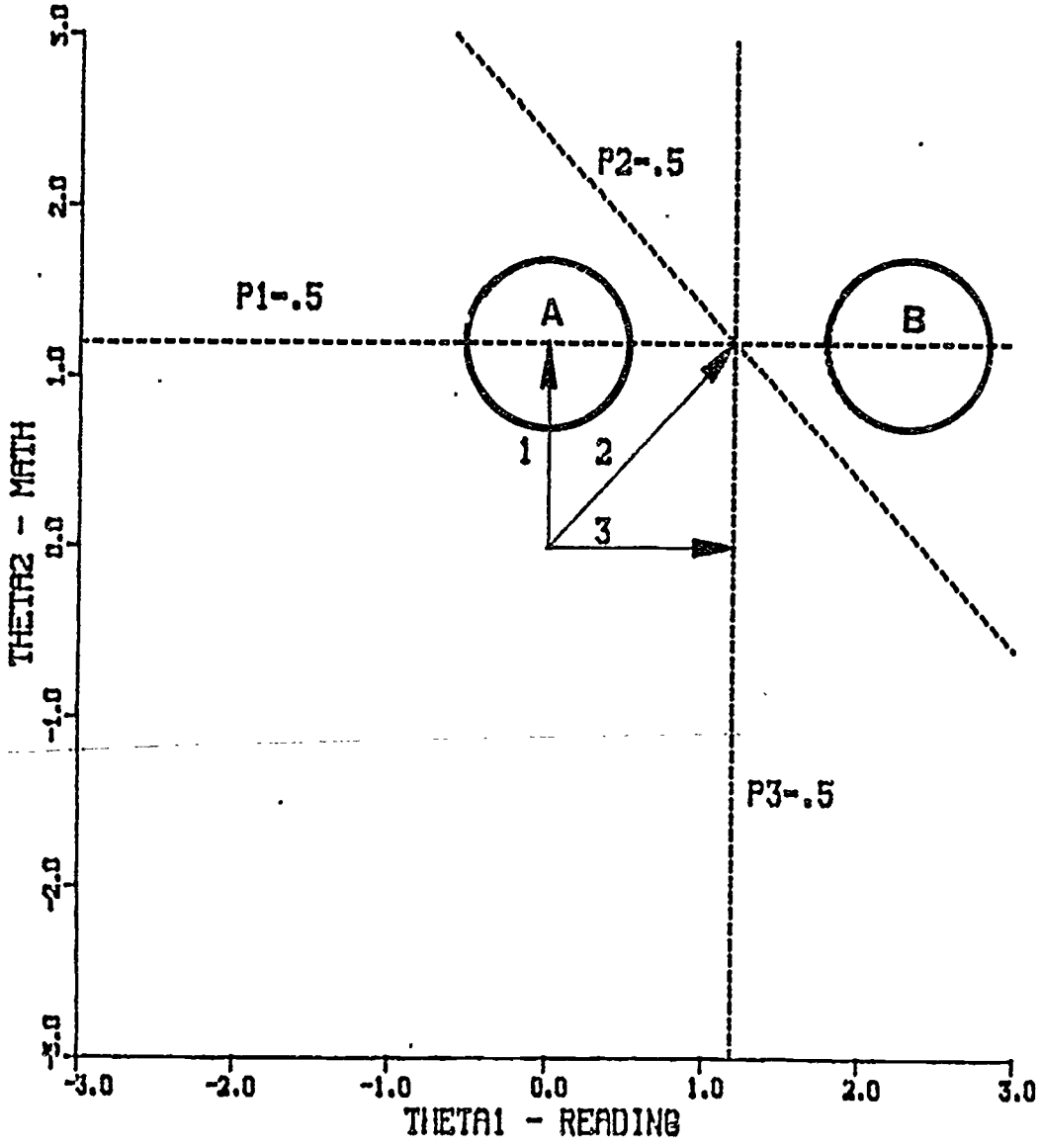
P2=.5
P1=.5
A
B
1
2
3
P3=.5
THETA2 - MATH
THETA1 - READING
3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0

FIGURE 2

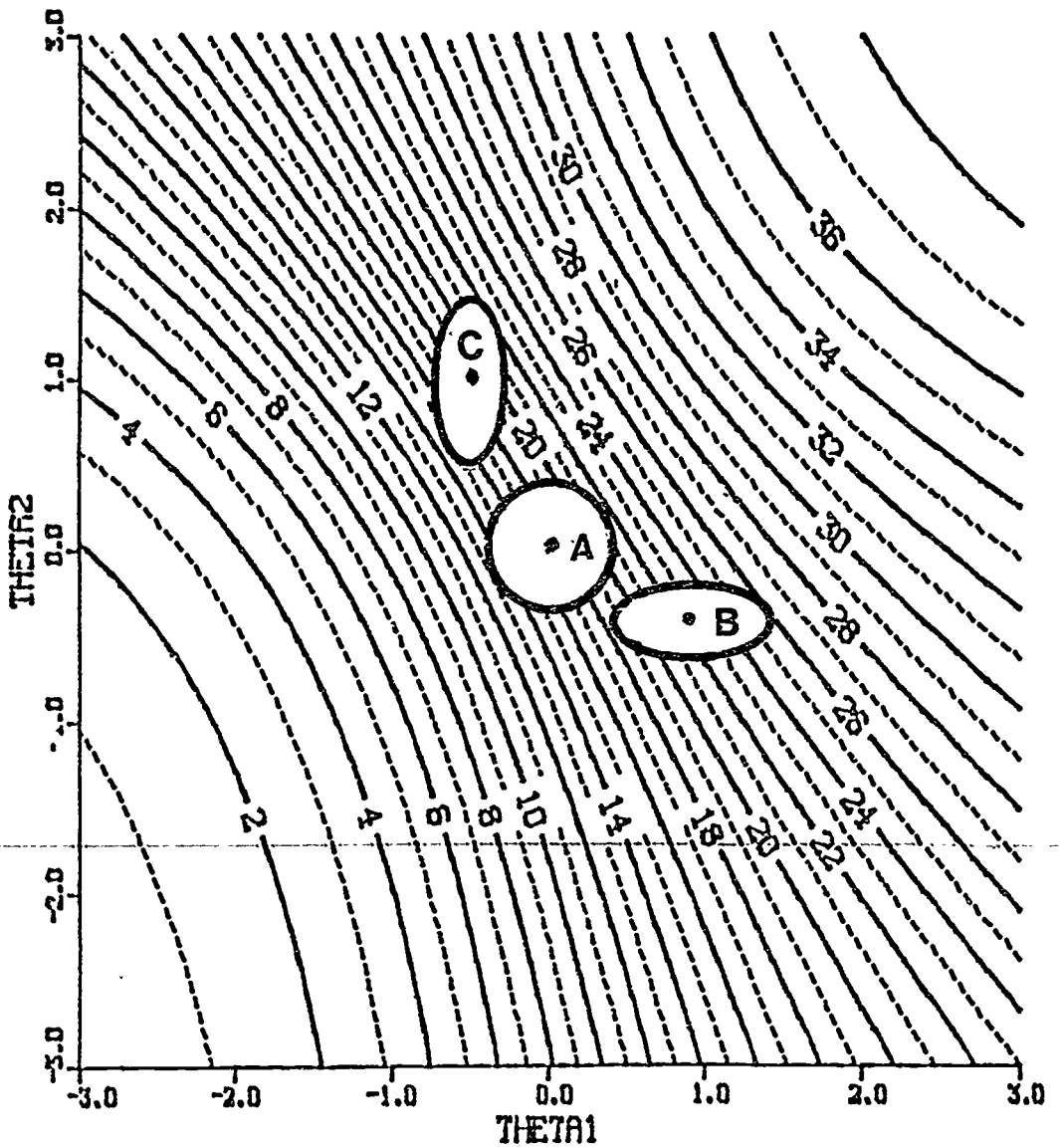

FIGURE 3

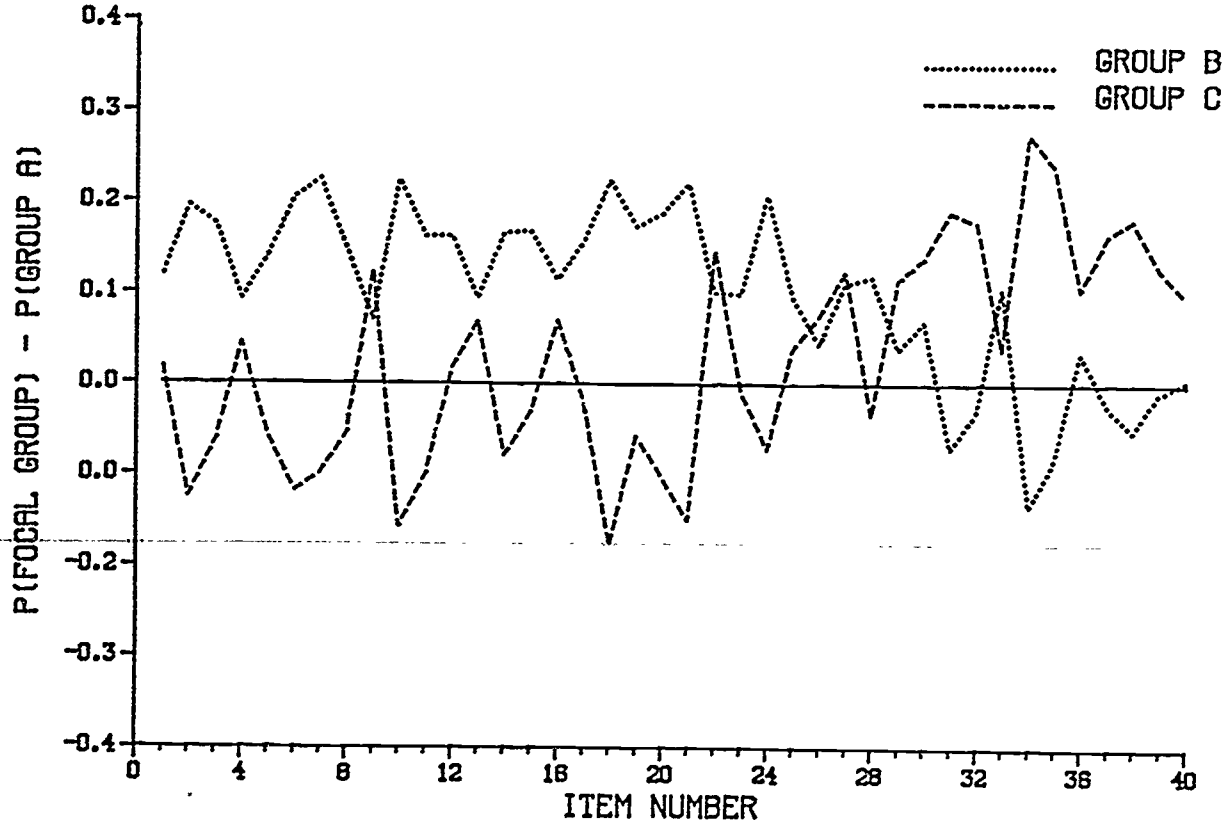
GROUP B
GROUP C
P(FOCAL GROUP) - P(GROUP A)
0.4
0.3
0.2
0.1
0.0
0.0
-0.2
-0.3
-0.4
0
4
8
12
16
20
24
28
32
36
40
ITEM NUMBER

FIGURE 4